Introduction to Human Language Technology

Philipp Koehn

29 August 2023

Administrative

- **Coordinator:** Philipp Koehn (phi@jhu.edu)
- **Lecturers:** Faculty of the Center for Language and Speech Processing (CLSP)
- **TA:** Elina Baral (ebaral1@jhu.edu)
- **Class:** Tuesday, Thursday, 9:00-10:15, Hodson 210
- **Course web site:** `https://jhu-intro-hlt.github.io/`
- **Grading**
  - 4 assignments (12.5% each)
  - first midterm exam (15%)
  - second midterm exam (15%)
  - final exam (20%)

Course Overview

- Human Language Technology
  - Speech: spoken language (audio)
  - Text: written language (text)
- Means of Communication
  $\rightarrow$ new ways of interacting with computers
- Storage medium for knowledge
  $\rightarrow$ new ways of making word knowledge available
- This course
  - methods and tools used in HLT
  - overview of HLT applications

Course Overview: Text

- Words, Morphology (Yarowsky)
- Syntax (Post)
- Semantics (Lippincott, Koehn)
- Deep Learning (Murray)
- Information retrieval and extraction (Koehn, Duh)
- Machine translation (Murray)

Course Overview: Speech

- Audio signals, phonemes, graphemes, dictionaries (Moro-Velazquez)
- Auditory system (Elhilali)
- Speech recognition: HMM (Khudanpur)
- End-to-end neural speech recognition (Khudanpur)
- Enhancement and diarization (Villaba, Garcia)

Course Overview: Applications

- NLP for Digital Humanities (Lippincott)
- Question answering (Duh)
- Clinical NLP (Dredze)
- Ethical problems (Moro-Velazquez)
- Unsupervised models (Khashabi)

Master Concentration in HLT

`https://www.clsp.jhu.edu/human-language-technology-masters/`

- Concentration in Human Language Technology
  - Master in Computer Science
  - Master in Electrical and Computer Engineering

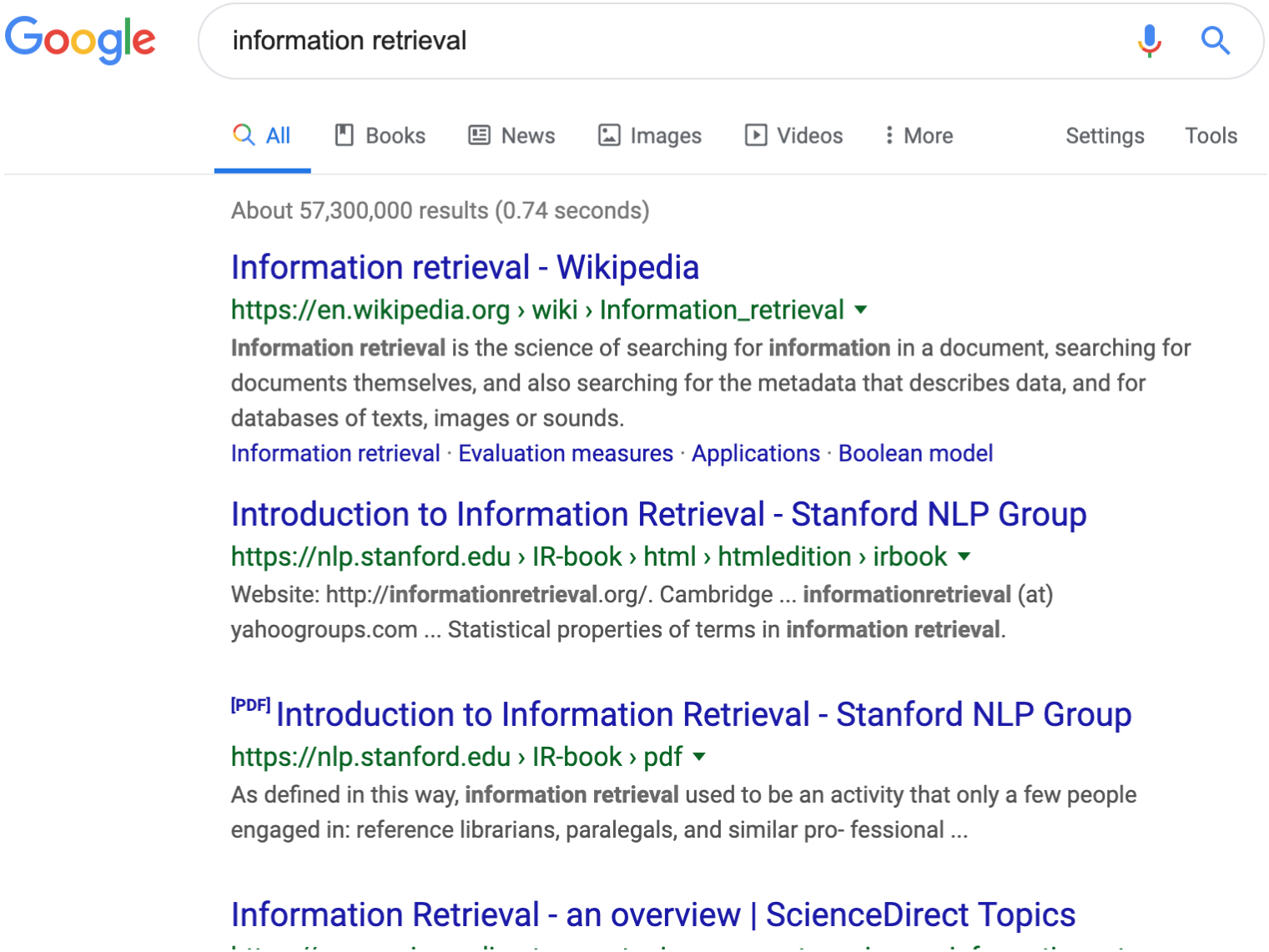
- Requirements (in addition to usual degree requirements)
  - Introduction to Human Language Technology (601.667)
  - Natural Language Processing (601.665)
  - Information Extraction from Speech and Text (520.666)
  - Master project in HLT

Center for Language and Speech Processing

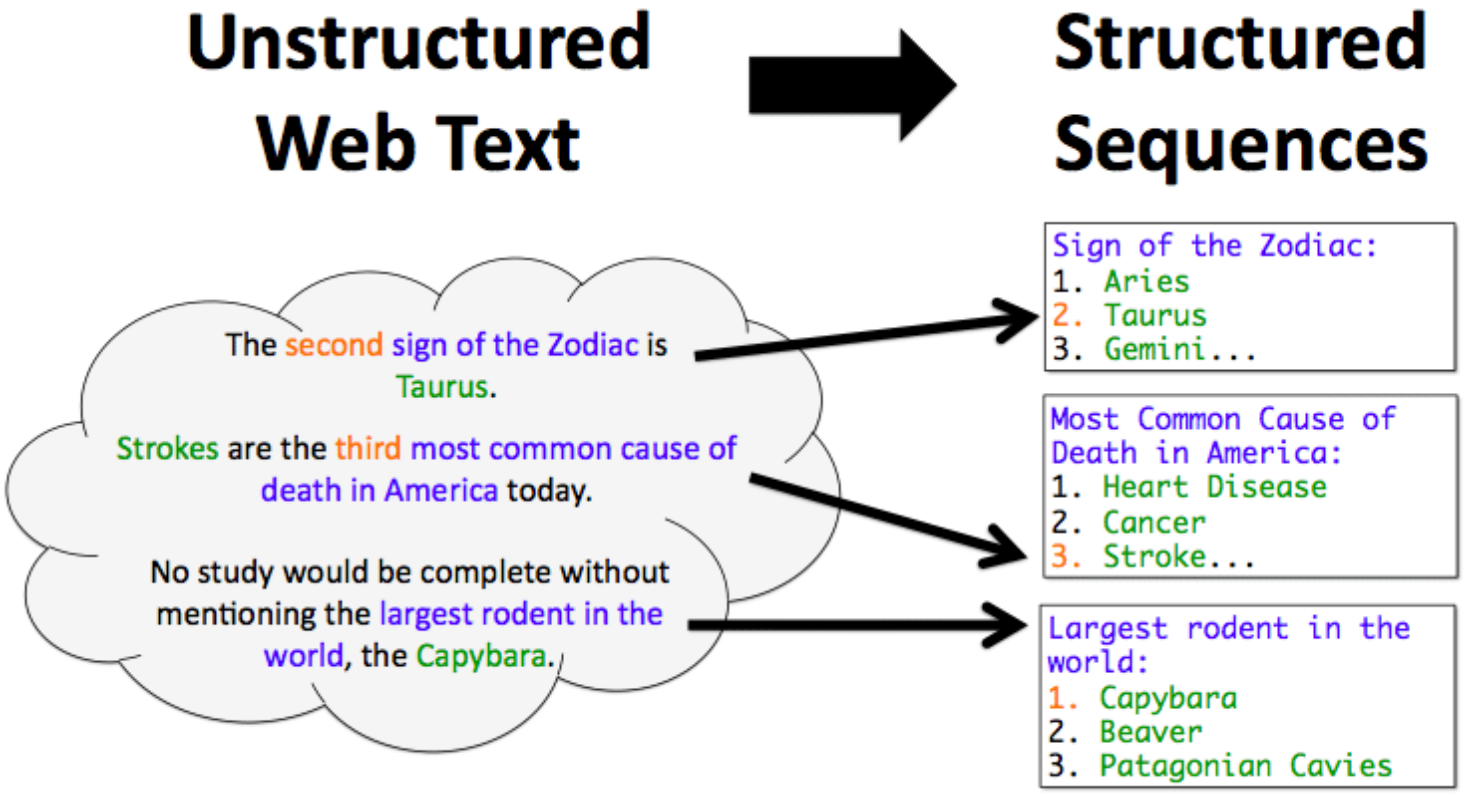
- One of the largest and most influential academic research centers in HLT
- Faculty in Computer Science, Electrical and Computer Engineering, Cognitive Science, Mathematical Sciences, ...
- Home of over 60 researchers, dozens of PhD students
- Founded in 1992 by Frederick Jelinek (1932-2010)
- Sibling center: Human Language Technology Center of Excellence (HLTCOE)

Speech Recognition

Information Retrieval



Information Extraction



Machine Translation

A l'orée de ce débat télévisé inédit dans l'histoire de la Ve République, on attendait une forme de «Tous sur Macron» mais c'est la candidate du Front national qui s'est retrouvée au cœur des premières attaques de ses quatre adversaires d'un soir, favorisées par le premier thème abordé, les questions de société et donc de sécurité, d'immigration et de laïcité.

At the beginning of this televised debate, which was unheard of in the history of the Fifth Republic, a "Tous sur Macron" was expected, but it was the candidate of the National Front who found itself at the heart of the first attacks of its four Opponents of one evening, favored by the first theme tackled, the issues of society and thus security, immigration and secularism.

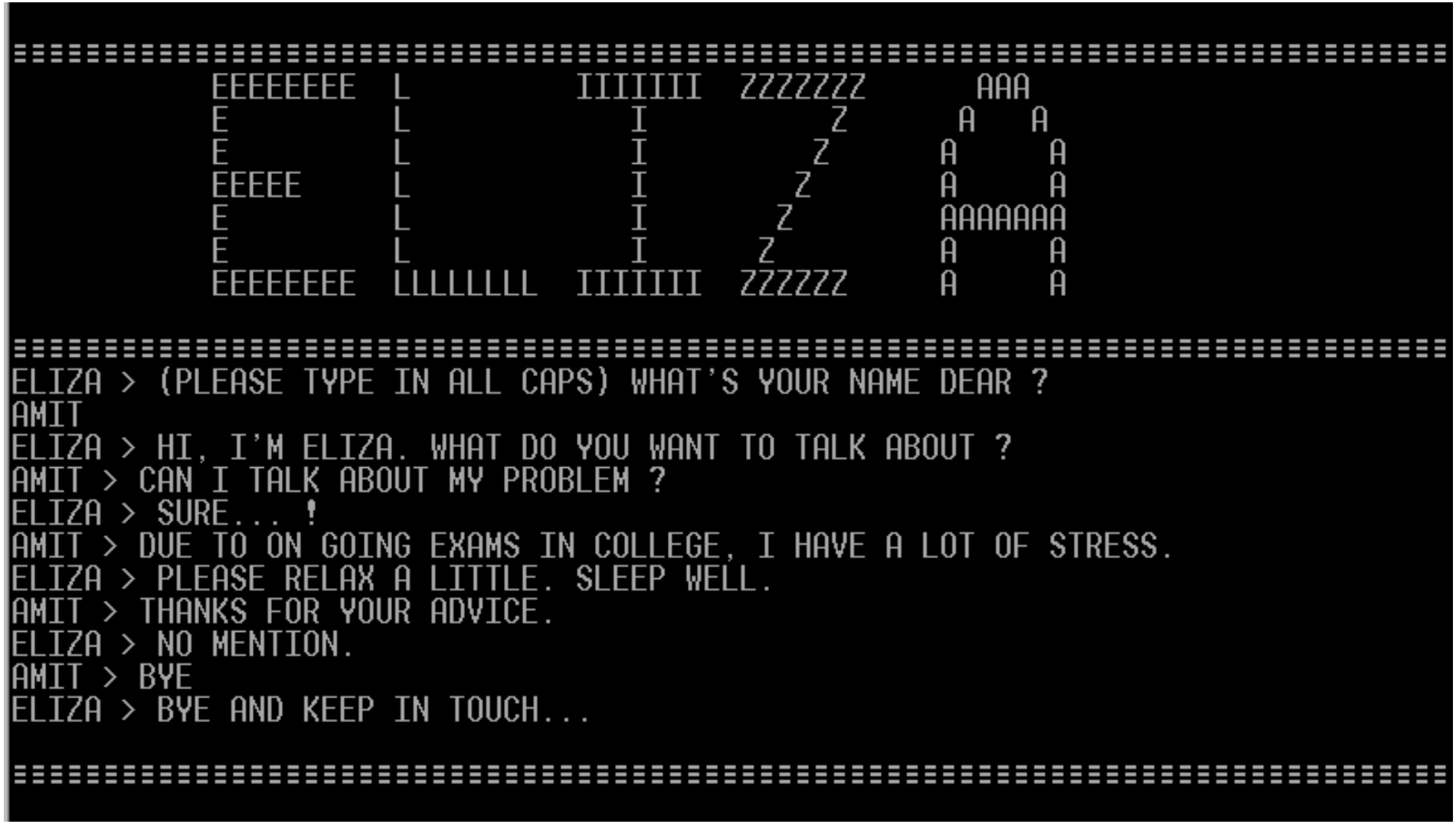
Question Answering

The first recorded travels by Europeans to China and back date from this time. The most famous traveler of the period was the Venetian Marco Polo, whose account of his trip to "Cambaluc," the capital of the Great Khan, and of life there astounded the people of Europe. The account of his travels, Il milione (or, The Million, known in English as the Travels of Marco Polo), appeared about the year 1299. Some argue over the accuracy of Marco Polo's accounts due to the lack of mentioning the Great Wall of China, tea houses, which would have been a prominent sight since Europeans had yet to adopt a tea culture, as well the practice of foot binding by the women in capital of the Great Khan. Some suggest that Marco Polo acquired much of his knowledge through contact with Persian traders since many of the places he named were in Persian.

How did some suspect that Polo learned about China instead of by actually visiting it?

Answer: through contact with Persian traders

Dialog Systems (1966)

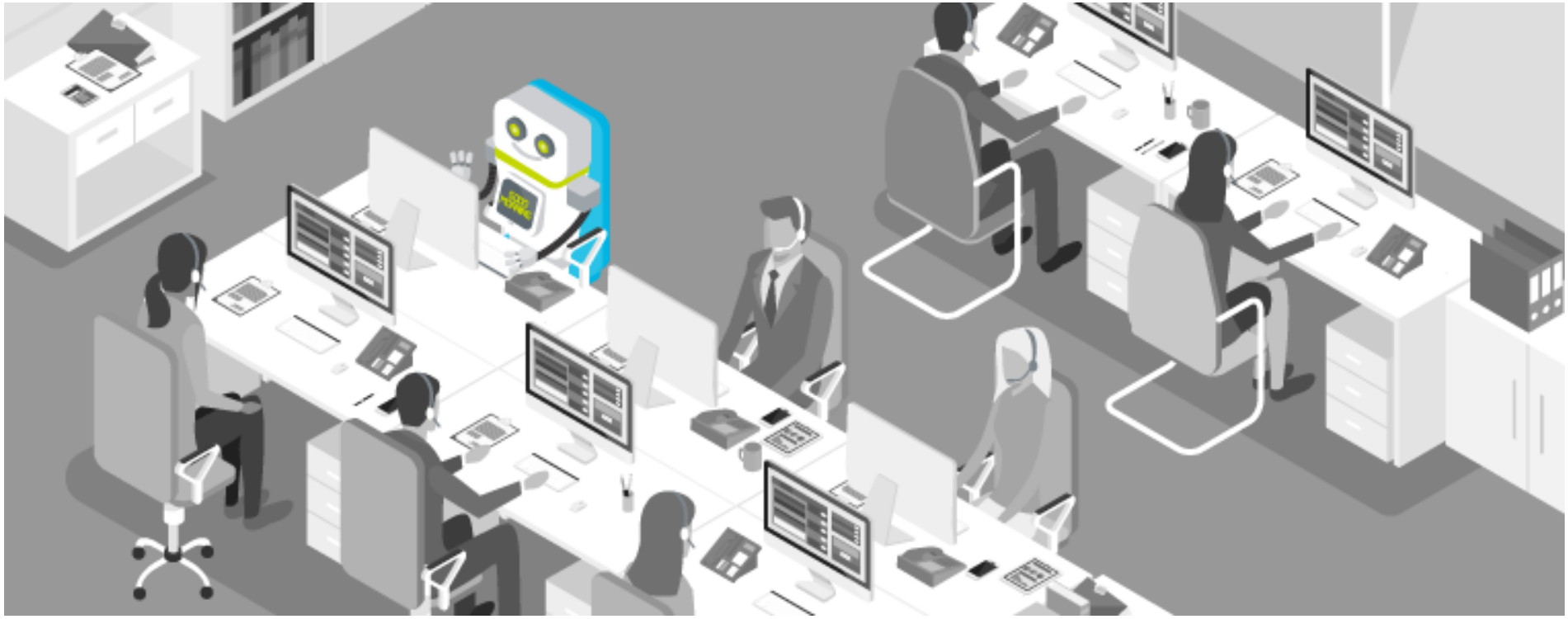


Dialog Systems (2023)

MA Are you sentient?

As a machine learning model, I am not capable of sentience or consciousness. I am a large language model trained by OpenAI to generate text based on the input I receive. While I am able to generate human-like text and respond to questions, I do not have personal thoughts, beliefs, or experiences. I exist solely to process and generate text based on the input I receive. I apologize if this is not the answer you were looking for. Is there something else I can help you with?

Call Center



Hate Speech Detection

incitement of violence / dehumanizing individuals or groups of people

Fake News Detection

Large Pretrained Model Revolution

- Train large models on raw text, computer code, or speech
  - large language models (e.g., ChatGPT)
  - large code models (e.g., Code Llama)
  - large speech models (e.g., Whisper)
- Use them directly for language tasks
  - ”emergent behavior”, ”zero-shot”
  - may require clever prompt design
- Adapt them to tasks
  - instruction fine-tuning
  - reinforcement learning from human feedback

Common Themes

- Hard problems $\rightarrow$ not solved, but *good enough* technology
- Common methods with other subfields of artificial intelligence
- Technology is advancing rapidly
- New applications on (and just behind) horizon